GENERAL SYSTEM FOR NORMAL AND PHONETIC INFLECTION

Stefan DIACONESCU[*], Cristi INGINERU[*], Felicia CODIRLASU[*], Monica RIZEA[*], Oana BULIBASA[*]
[*] Department of Research and Development
SOFTWIN
Bucharest, Romania
sdiaconescu@softwin.ro

Abstract— Generating all inflected forms for a natural language is a very difficult task not only because of the large number of inflection rules but also because of the large number of exceptions. The operation is more difficult if we consider not only the normal inflected forms (where the inflected forms are represented in the normal alphabet of the involved language) but also the phonetic forms (where the inflected forms are represented in a phonetic alphabet). This paper presents a general system (i.e. a system that can be used for any inflected natural language) that generates all the normal and phonetic forms of words from a given lexicon. A set of results obtained by using this system for Romanian language is also presented. The system is based on GRAALAN metalanguage (Grammar Abstract Language) used for representing the linguistic knowledge concerning a natural language and requested by the inflection process. It also uses a set of tools that allows the handling of this (meta) language. The system starts with a set of linguistic knowledge bases described in GRAALAN containing: normal and phonetic inflection rules, the description of the normal and phonetic alphabets that are used, the base forms (lemmas) that will be inflected, taken from a lexicon (also in normal and phonetic forms) and the description of the morphological structure of the language (morphological categories and their values). All these linguistic knowledge bases are created using special tools. The inflected normal/phonetic forms will be automatically generated by applying the normal/phonetic inflection rules on the lemmas of the lexicon, taking into account the morphological structure and the normal/phonetic alphabets. At the end of the process, the linguist is able to verify and eventually correct the generated forms. When corrections are made, new inflection rules are automatically generated. The presented system is part of larger system for natural language processing.

Keywords-component; inflection rules; phonetic inflection rules; inflected forms; natural language processing

I. INTRODUCTION (HEADING 1)

Though it seems to be quite peculiar, there is not a generally accepted definition of the "word" notion. According to [10], the word is "the basic unit of the vocabulary that represents the association of a sense (or of a complex of senses) and a sound complex". In [15] it is given a more general definition ("meaningful unit of spoken language that can stand alone as utterance and is not divisible into units") and also a more restrictive one ("written or printed representation of a spoken word that is usually set off by spaces on either side"). In [3], there is a general definition ("the complex linguistic unit, simultaneous realised as phonetic, semantic and grammatical unit") and one that refers to written words ("characters group in an alphabetic, syllabic, ideograph transcription, set off by typographical white spaces"). In the general case, different structures of a word must be considered: phonological, morphologic, semantic and syntactic structures. All these definitions seem to be more appropriate for indo-european languages. The notion of "word" must be more refined for agglutinant or insulated languages. Consequently, the notion of "word" is finally dependent on the considered natural language. By a typological classification, there are two types of natural languages: insulated (like Chinese) or non insulated. In the insulated languages, words are generally invariable, the different situations of a word being represented by syntactic elements (that consider the word order for example) and intonation. The non insulated languages can be agglutinant or inflected. In agglutinant languages (like Turkish), words are formed by a root and a number of affixes. Sometimes it is not very easy to make the difference between agglutinant and inflected characteristics. The inflected languages can be grouped into: synthetic and analytic languages. In synthetic languages words are formed by many morphemes that fuse together. The analytic languages use syntactic means in order to obtain some situations of words. Actually, there is not a language of pure type, the characteristics of different types being combined at different degrees. For example, Romanian and French languages combine synthetic and analytic features, yet Romanian language is more synthetic than French language.

We will consider here that a word is the basic unit of vocabulary and represents an association between a sense and a sign (or a string of signs). The sign can be a sound sign, graphic sign or any other type of sign.

GRAALAN system [8] implements GRAALAN language (Grammar Abstract Language) [7] and its purpose is to allow the description of certain linguistic aspects regarding one natural language or the correspondences between two natural languages. In GRAALAN, we will consider the graphic signs (coded according to UNICODE [12]) and the sound signs, which are also represented by some graphic signs, (according to IPA - International Phonetic Alphabet [10], for example, that is also UNICODE codified).

GRAALAN system is more appropriate for the non insulated inflected (synthetic or analytic) languages but it can also be used for other types of languages.

When a word is used in a phrase, it can be in different inflection situations according to the syntactic role that the word plays in the phrase. In GRAALAN, an inflected situation is an ensemble of morphological attributes and values of these attributes, organised as an AVT (Attribute Value Tree) [6]. An inflected form corresponds to each inflection situation. Many inflection situations can have the same inflected form (in some cases, for non inflected words or for insulated languages, the inflected form is unique therefore it loses attribute "inflected").

The set of all inflection situations of a word is named in GRAALAN word extended paradigm. Usually, a word paradigm is defined, in a more restrictive way, as the set of inflected forms. The larger definition in GRAALAN allows considering the non insulated languages using the same formalism.

Consequently, even languages with very poor inflection (like English) or (almost) no inflection (like Chinese) can be characterized by a quite large (extended) word paradigm.

Word inflection is represented in GRAALAN as the process of obtaining all the inflection situations of a word together with all the corresponding written inflected forms. This process refers not only to the normal alphabet written word (we understand here by alphabet the usual alphabet, and also syllabic or ideographic transcription) but also to the phonetic alphabet written word.

Unlike the other automatic inflection systems [2] [14], automatic phonetisation systems [1], and automatic syllabification systems [11], that usually produce different kinds of dictionaries specific to a certain language and function, GRAALAN system is a unitary system that groups all the above functions and many others, and gives them a structure as part of a larger NLP (Natural Languages Processing) system.

Section 2 will present how an inflected process takes place in a GRAALAN based processing system. The inflected process uses some previously created linguistic knowledge bases: morphological configurator (section 3), alphabet (section 4), lexicon (section 5), syllabification rules (section 6), and inflection rules (section 7). The inflection situations and the inflected forms that are generated during the inflection process are presented in section 8. The application MKT (Morphological Knowledge Tool), that accomplishes the inflected process itself, is presented in section 9. For each of sections 3 to 8 some effective implementation for Romanian language is presented too. In section 10, some conclusions about the generality of GRAALAN system/language and the usability in different linguistic applications are presented.

This template, modified in MS Word 2003 and saved as “Word 97-2003 & 6.0/95 – RTF” for the PC, provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. All standard paper components have been specified for three reasons: (1) ease of use when formatting individual papers, (2) automatic compliance to electronic requirements that facilitate the concurrent or later production of electronic products, and (3) conformity of style throughout a conference proceedings. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

II. THE PLACE OF INFLECTION SYSTEM IN A NLP SYSTEM

GRAALAN language allows the description of the following linguistic sections: alphabet (in a very large sense), morphological configurator, syllabification rules, lexicon, inflection rules, inflection forms, syntax rules, correspondence between two natural languages. The inflection process uses only the following sections: alphabet, morphological configurator, syllabification rules, lexicon, inflection rules, and inflection forms (see Fig. 1). (Actually, the syllabification does not belong to inflection process but it is completed in the same time and with the same tool as the inflection.)

Each of the five input GRAALAN sections must be described by the linguist for a certain natural language (see Fig. 2), using GRAALAN language (and tools). The GRAALAN text is compiled using a GRAALAN compiler and transformed in a knowledge base (LKB) corresponding to each section which is stored in XML (Extensible Markup Language) format [13]. The lexicon section can be also introduced using a special tool LKT (Lexicon Knowledge Tool) that automatically generates GRAALAN text. The lexicon GRAALAN text is compiled and, then, the lexicon XML LKB is generated.

Afterwards, the five XML LKBs are used by MKT, which implements the inflected normal and phonetic process and generates inflected forms in GRAALAN. Then, the inflected forms are compiled and transformed in a XML LKB (Fig. 1).

MKT is a tool that allows the linguist to also verify and to correct the inflected forms using an interactive interface. Should some modifications be necessary, (caused, for example, by some exceptions not yet considered), MKT will automatically generate in GRAALAN new inflection rules in order to specify this modifications. The new rules will be compiled and the XML LKB with inflected forms is updated.

The knowledge bases containing inflected forms will be then used by different tools and applications, as showed in Fig. 2.

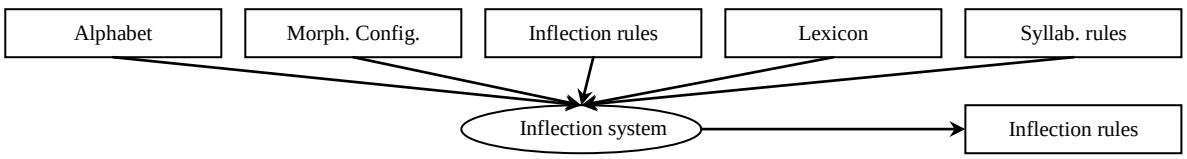

Figure 1. GRAALAN Sections involved in Inflection System

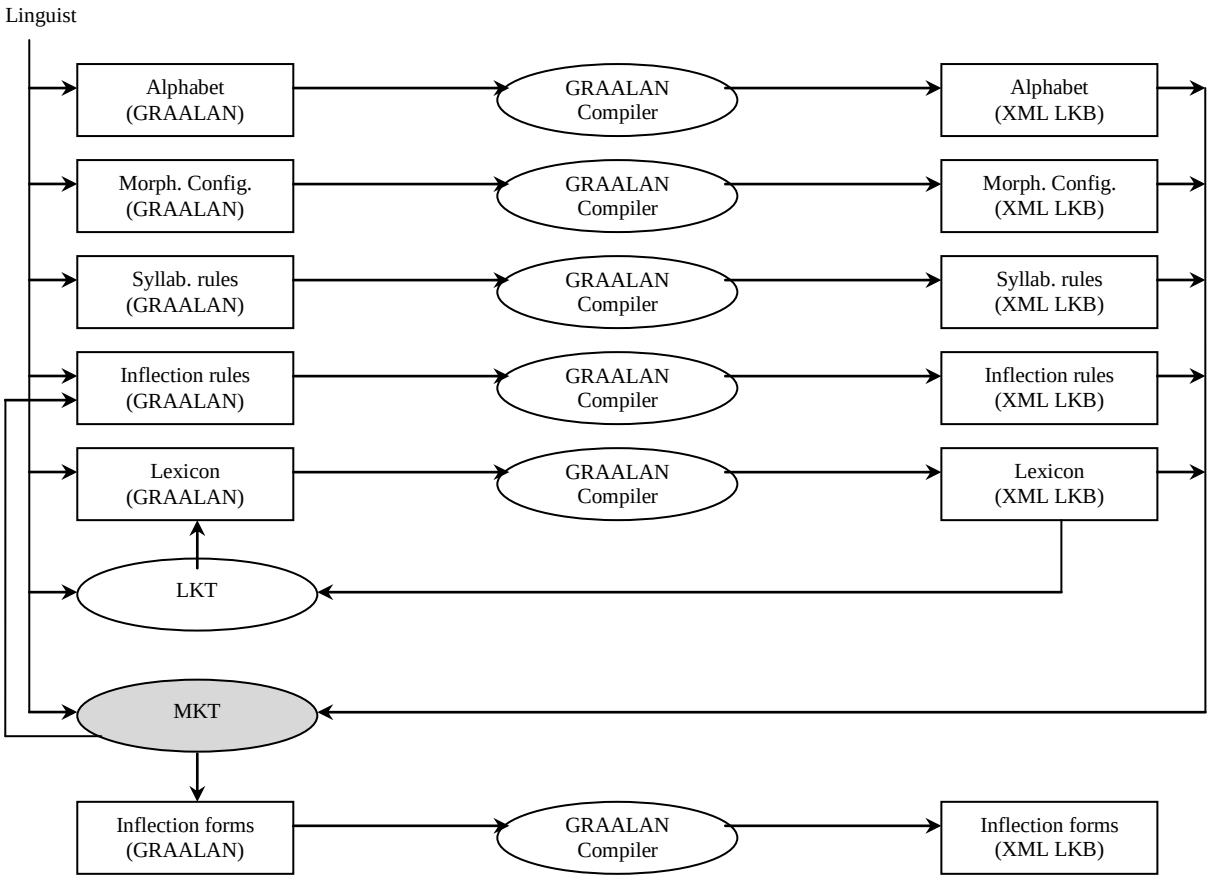

Figure 2. Inflection process in a complex NLP System

III. THE MORPHOLOGICAL CONFIGURATOR

The morphological configurator of a natural language represents the ensemble of morphological categories (e.g. class, gender, number, case, etc.) and values of these morphological categories (for example, noun, adjective, verb, etc. for class; masculine, feminine, neuter for gender; plural, singular for number, etc.) organised under the form of an AVT [6], where the attributes correspond to the morphological categories and the attribute values correspond to the morphological attribute values.

An AVT has as nodes attributes and attribute values. We can define (very simplified) the structure of an AVT (in BNF notation):

1. <AVT> -> [<attribute> = <attribute value list>]
2. <attribute value list> -> <attribute value element>, <attribute value list> | <attribute value element>
3. <attribute value element> -> <attribute value> <AVT> | <attribute value>

A path in an AVT is formed by a sequence of [<attribute> = <attribute value>] pairs. Such a sequence is named EC (Exclusive Combination). An EC represents an inflection situation.

The (expanded) paradigm of a word covers a number of inflection situations form the morphological configurator, i.e. a number of ECs. Therefore, the morphological configurator represents the set of all the inflection situations that all the words (from the lexicon) can have. For a not inflected word, we will have only one inflection situation.

In Fig. 3 an extremely simplified AVT is represented. It could also be represented as follows:

[class = noun [gender = masculine, feminine, neuter]
[number = singular, plural]]

It will result eight ECs and many inflection situations:

[class = noun] [gender = masculine] [number = singular]

[class = noun] [gender = masculine] [number = plural]

[class = noun] [gender = feminine] [number = singular]

[class = noun] [gender = feminine] [number = plural]

[class = noun] [gender = neuter] [number = singular]

[class = noun] [gender = neuter] [number = plural]

[class = verb] [number = plural]

[class = verb] [number = singular]

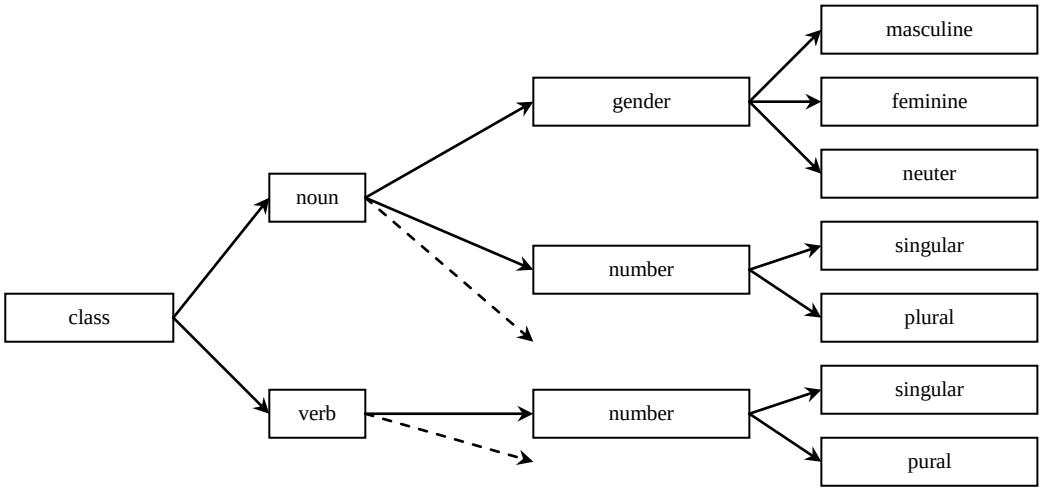

Figure 3. Example of a simplified AVT

In GRAALAN, each node has associated not only the attribute name or the attribute value name, but also some other information:

a) *For attributes*: the abbreviation of the name, the indication if the morphological category is inflected or not, the name of a function that procedurally makes the treatment, if it is the case. (GRAALAN is a declarative language but it can call some procedural function, if it is needed.)

b) *For attribute values*: the abbreviation of the name, the indication if the value is applied to the lemma, the name of a function that procedurally makes the treatment, if it is the case, etc.

For the morphological configuration implemented for Romanian language, we established a variant with (see Table I): 11 basic classes (noun, article, pronoun, adjective, numeral, verb, adverb, preposition, conjunction, interjection, and a special class named "sign"), 819 attribute apparitions (among these there are 111 inflected attribute apparitions) using 67 distinct attribute names, and 2,122 attribute values apparitions using 211 distinct attribute value names. This morphological configurator defines 19,462 ECs (the verb, of course, has the most ECs: 16,320). The maximum EC length is 36.

We must say that the description in a morphological configurator is not unique. It can be accomplished using more or less attributes and attribute values, according to the target of the research. This implies that the number of inflected situations will be different depending on the number of attributes and attribute values that are used.

TABLE I. ROMANIAN MORPHOLOGICAL CONFIGURATOR

Name	Total attributes	Inflected attributes	Total values	EC Number	Maximum EC length	Attributes names	Values names
Class	819	111	2,122	19,462	36	67	211
Class = Noun	9	3	27	312	12	6	18
Class = Article	13	2	38	82	8	4	15
Class = Adjective	12	0	30	420	14	9	21
Class = Pronoun	112	12	324	1,118	16	13	45
Class = Numeral	154	43	447	1,124	16	20	63
Class = Verb	478	35	1,142	16,320	34	24	73
Class = Adverb	33	8	89	68	14	12	32
Class = Preposition	1	1	4	3	2	1	4
Class = Conjunction	3	3	10	7	4	3	10
Class = Interjection	2	2	5	3	4	2	5

IV. THE ALPHABET

In GRAALAN, by the alphabet of a language, we understand the set of signs (usual alphabetic signs, syllabic signs, and ideographic signs) with their codes that are used to write texts and to describe phonetic aspects of that language. In a GRAALAN alphabet section, the following subsections are presented:

a) *The normal alphabet* is the set of signs used to write texts in the language.

b) *The phonetic alphabet* is a set of signs used to phonetic transcription of the language sounds.

c) *The special characters* are other signs used, for example, punctuation marks.

d) *Stress markers* specify the primary stress and the secondary stress.

e) *Character groups* (like diphthongs and trifthongs for Romanian language) indicate the relation between the normal written signs and the phonetic written signs.

f) *Classes* define the names for different categories of signs (for example, the consonant class, the vowel class, the capital letters, etc.).

Table II shows the number of alphabet signs used for Romanian language in the present implementation.

TABLE II. GRAALAN SIGNS FOR ROMANIAN LANGUAGE

No.	Sign types	Signs number	Examples of sign names
1	***normal alphabet***	66	A, a, B, b, ...
2	***phonetic alphabet***	36	open_central_unrounded, mid_central_unrounded, mid_front_unrounded
3	***special characters***	64	tilde, grave_accent, commercial_at
4	***stress markers***	2	primary_stress, secondary_stress
5	***groups***	360	a_group, tch_group, sh_group
6	***classes***	17	capital_letter, small_letter, vowel, semivowel, consonant, diphthong

Additional information for different types of characters can be specified: codes, names, sorting sequence, direction of writing (to the left, to the right), some special functions, etc.

V. THE LEXICON

GRAALAN lexicon is a very complex data structure, much more complex than a usual lexicon. We will mention here only few of the lexicon features, those that are involved in the inflection process. A GRAALAN lexicon is a section with a large number of entries. There are three types of entries: lexical entries, morphological entries end procedural entries.

a) *Lexical entries* contain morphemes, words and multiword expressions (MWE). Morphemes are particles that contribute to word formation: roots, prefixes, suffixes, prefixoids, suffixoids, prime elements, median elements, final elements. Words are the basic elements of the syntactic constructs: lemmas (basic or canonical word forms – for example, the nominative, singular forms for nouns), supplements (for example, the nominative plural forms for nouns), different other word forms (from the word paradigm, for example, the second person of the personal noun). Multiword expressions are groups of words that have a global sense, very different from the senses of each single word. In GRAALAN lexicon, these MWEs are represented not only as a word sequence but in a more elaborated form (using a dependency tree DT).

b) *Morphological entries* contain information about the language morphology and are correlated with morphological configurator: morphological categories and values, synthetic characteristics (for the mono-word inflected forms) and analytical characteristics (for the multiword inflected forms).

c) *Procedural entries* describe the procedures that are eventually called in different parts of the linguistic description.

From the inflection point of view, the most important type of entry is the lexical lemma. A lemma is in fact, in GRAALAN, one of the inflected forms associated to an inflection situation from the extended paradigm of a word, used in a conventional mode as a reference form for the others inflected forms. A lemma can eventually contain more words (for example, "a zbura" – to fly – in Romanian language). One of the words of such a multiword lemma is considered as inflected form's center and it is used in the sorting process. The other words of the inflected form are considered auxiliary words. Each auxiliary word will be accompanied by different information: references to lemma the auxiliary word originates from, an associated AVT. Consequently, a multiword lemma entry in the lexicon is represented as a DT structure.

Among the information that represents an entry lemma in a GRAALAN lexicon, we find the following information involved in inflection process: the lemma itself in normal and phonetic alphabet, the lemma syllabification in normal and phonetic alphabet, (eventually) the morphological syllabification, the lemma AVT, the structure (DT) for the multiword lemmas, and the inflection rule label that must be applied to the lemma in order to obtain all lemma extended paradigm.

The current GRAALAN Romanian language lexicon (not yet finished) contains: about 76,000 lemmas, (66,000 mono-word lemma, 10,000 multiword lemmas), 115,000 senses, 103,000 supplements, and 12,700 multiword expressions.

VI. THE SYLLABIFICATION RULES

The inflected process in GRAALAN system is accomplished in parallel with the inflected form syllabification. The rules for the syllabification are described in a special GRAALAN section. There are three types of syllabification considered by GRAALAN:

a) *The normal (or euphonic) syllabification* that refers to normal alphabet written words.

b) *The phonetic syllabification* that refers to the phonetic alphabet written words.

c) *The morphological syllabification* that refers to the normal alphabet written words but also considers some morphological criteria [9].

In fact, only the first two syllabifications have syllabification rules and these rules must be described in the corresponding GRAALAN section.

An example of normal syllabification is the following:

Rule "&vowel;" + "&semivowel;" - "&semivowel;" + "&semivowel;" + "&vowel;" ;

By this notation, the sign "+" means "must not be separated" and the sign "-" means "it can be separated". The "&vowel;", "&semivowel;", "&semivowel;", "&semivowel;", "&vowel;" noted characters (according to XML conventions) refer to normal characters.

The corresponding sample for phonetic syllabification is the following:

Rule "&phon_vowel;"+"&phon_semivowel;" - "&phon_semivowel;" + "&phon_semivowel;" + "&phon_vowel;" ;

The "&phon_vowel;", "&phon_semivowel;", "&phon_semivowel;", "&phon_semivowel;", "&phon_vowel;" noted characters entities refer to phonetic characters.

The current GRAALAN syllabification implementation for Romanian language contains 723 normal syllabification rules and the same number of phonetic syllabification rules.

VII. THE INFLECTION RULES

GRAALAN collection of inflection rules for a certain natural language contains the rules used to obtain all the inflection situations of the lemmas from the lexicon and also the associated inflected forms represented in normal alphabet and phonetic alphabet.

The inflection rules are organised on four levels: compound rules, basic rules, elementary rules and variants (see Fig. 4). Each lemma from the lexicon indicates a compound inflection rule that is applied in order to obtain the extended word paradigm. A compound inflection rule is a list of basic inflection rules. A basic inflection rule is formed by a set of inflection situations that can be generated by the current basic inflection rule (represented as an AVT) and a set of elementary inflection rules. An elementary inflection rule contains a reference (to a lemma or to another inflection situation) and a list of variants. The transformations needed by the inflection process will be executed on the inflected form associated to this referred inflection situation. One variant (from the elementary inflection rule) contains a condition, a set of normal transformations, a set of phonetic transformations, and a structure (DT and AVT) for multiword inflection forms. The condition is an expression with a logical value (yes / no) that has as operands lemmas, or other elements from the current word paradigm, normal or phonetic character strings, groups or classes (see alphabet section), and as operators the identity operators. The normal transformation is an operation (insert, replace, or delete) executed on normal representation of the inflection form associated to the referred inflection situation. The phonetic transformation is an operation (insert, replace, or delete) executed on phonetic representation of the inflection form associated to the referred inflection situation.

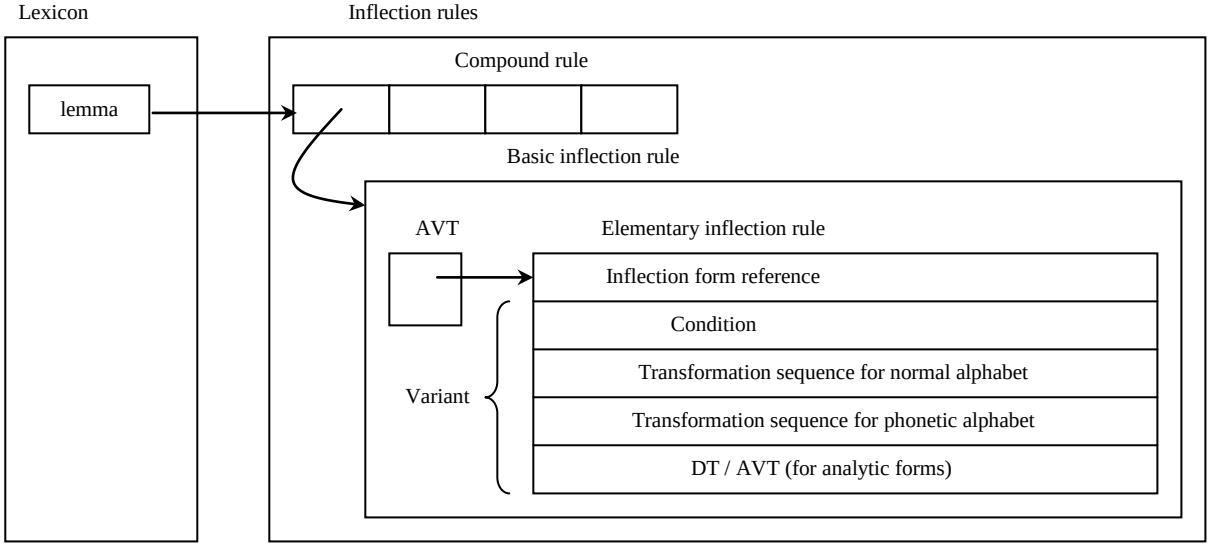

Figure 4. Inflection rules

A fragment of a basic inflection rule for Romanian language is the following:

Basic Rule Subst_masc1:

[clasa = substantiv] [tip substantiv = comun] [animatie = animat, inanimat] [gen = masculin]

[numar = singular [caz = nominativ

[articulare = nearticulat (EtL1: alphabetic - phonetic -)

, hotarat (EtS11:

if(&consonant;) alphabetic insert "ul" /* last letter is a consonant - băiat, elev */

phonetic insert "&close_back_rounded;&alveolar_lateral_approximant;"

if("i") alphabetic insert "ul" /* last letter is "i" - ochi */

phonetic insert "&close_back_rounded;&alveolar_lateral_approximant;"

…]

, genitiv … , dativ … , acuzativ …]

, plural …]]

We give below some figures in order to have an idea on the complexity of inflection rules for Romanian language:

Inflection situations (EC) that have inflection rules: 19,202

Inflection situations that do not have inflected rules: 260 (these situations are procedural treated, for example, some complex numerals)

- Variants: 28,317
- Multiword variants: 19,935
- "Defective" situations: 2,936
- Mono-word variants: 8,382
- Multiword variants with 2 words: 7,785
- Multiword variants with 3 words: 6,554
- Multiword variants with 4 words: 3,196
- Multiword variants with 5 words: 1,908
- Multiword variants with 6 words: 492.

VIII. THE INFLECTION FORMS

The inflection forms are obtained by applying the inflection rules on lemmas from the lexicon. A collection of inflection forms is a set of entries. An entry corresponds to an inflected form. An inflected form can have one or many words. When an inflected form has many words, one of these words is considered center and it will be used to do eventually an inflection forms tri. An entry contains:

- The characterisation of the inflection situation/situations of the inflected form (i.e. an AVT). Therefore, we group more inflection situations that have associated the same inflected form.

- The inflection form in normal and in phonetic alphabet.

- The label of the word from the lexicon where we find the lemma of the current inflection form.

- The normal, phonetic, and, eventually, the morphologic inflection form and the syllabification at the end of a line (the "hyphenation").

- The indication of the word used for tri (in the case of multiword inflection forms).

- The structure (DT) in the multiword inflected form.

A sample of an entry in inflection form's list for Romanian language is the following:

ETF_Entry_1-1-1_1_1: *Entry* Text "un actor" Phonetic "'un akt'or" Reference Entry_1-1-1_1

[clasa = substantiv] [tip substantiv = comun] [animatie = animat]

[gen = masculin] [numar = singular] [caz = nominativ, acuzativ] [articulare = nehotarat]

Syllabification Euphonic "un ac"-"tor" Phonetic "'un ak"."t'or" Tri 1 left

Central word Text "actor" Phonetic "akt'or"

[articulare = nearticulat] [numar = singular] [caz = nominativ] [gen = masculin] [animatie = animat]

[tip substantiv = comun] [clasa = substantiv]

Auxiliary words

Text "un" Phonetic "'un" Reference Art01

[clasa = articol] [tip articol = nehotarat] [caz = nominativ] [gen = masculin] [numar = singular]

Belongs = yes @acord-art@

end of entry

The current implementation for Romanian language generates: 17,928,056 inflection situations (2,075,978 mono-words and 15,852,078 multiwords) and 9,946,686 inflection forms (1,019,783 mono-words and 8,926,903 multiwords).

IX. INFLECTION EXECUTION

The inflection process is realised by a special tool named MKT. MKT had to be designed not only to automatically generate the inflection situations / inflected forms but to help the linguist to correct inflected forms of a particular lemma, to save and to generate new inflection rules according to changes made by the linguist. The system should allow also the linguist to check the inflected forms previously saved.

The actions that a linguist can apply to generated inflection forms are: editing, deleting, and inserting. Other possible actions are creating syllabification forms and choosing values for not inflectional attributes for the purpose of generating the entire inflection situation tree for the current lemma. In order to decrease the entire correction of inflection forms time, the system generates and displays those inflection forms of more than one lemma in the same time. This facility saves some dead times of loading rules, also saving some other various actions that the linguist can take for all inflected forms of more than one lemma.

The only condition for simultaneously working with more lemmas is that all lemmas must have the same inflection situation and the same compound rule. All the inflection situations with the corresponding inflected forms of a particular lemma constitute the word tree (see Fig. 5). The word tree is obtained from morphologic configurator filtered on the

inflection situation of a lemma. The filtering process consists in deleting the undesired values of not inflectional attributes according to the value of attributes from the lemma's inflection situation. For example, if the lemma's inflection situation says that it is about a verb then all other lexical classes are ignored. A basic inflection rule is valid only in a special context when it is included in a compound rule beside other basic rules because all elementary rule labels must be unique and all references must be solved. The program had been limited to use only an "active" compound rule at a given time. Active compound rule means that all its refered basic rules are found in memory and the elementary rule label references have been searched and indexed for a fast access during runtime. In conclusion, from the MKT point of view, lemmas can be grouped by inflection situation and inflection rule.

There are two kinds of steps for generating inflection forms: the word tree preparation and the inflecting itself.

I. *The word tree preparation*. It is made only once, after the linguist has been selected the lemma. It works for more lemmas in the same way that it works for only one. The input of this section is the inflection situation and the morphologic configurator. The result is the word tree which later will be filled with inflected forms.

a) *Copying morphologic configurator* – a new copy of morphologic configurator is saved and the original one is kept for further usages.

b) *Loading lemma's inflection situation* – the system loads the inflection situation of current lemma or lemmas from the lexicon.

c) *Intersection* – the morphologic configurator copy is filtered using the inflection situation of the lemmas from the lexicon. The result is the word tree.

d) *Expanding* – the word tree is being expanded which means that the attributes will have each at most one value.

II. *The inflecting*. During a work session, the user can modify some automatic generated inflected forms. He can, eventually cancel the modifications made by himself and command the new automatic inflected form generation, etc. The following steps are selected according to the user actions.

a) *Loading inflection rule* – this step involves the reading of the inflection compound rule and the associated basic inflection rules.

b) *Loading inflection forms* – only if those were generated and saved before.

c) *Selecting values for not inflectional attributes* – if the current lemma's inflection situation is not enough to build the exact word tree, the inflection situations of saved inflection forms are being used to choose values for not inflectional attributes. If there were not previously saved inflection forms, some information has to be supplied by the linguist using the graphic interface.

d) *Generating inflection form nodes* – for all inflection situations from the word tree, the system searches for the corresponding elementary inflection rules. For all elementary rules attached to the inflection situation found in basic rules, new inflection form nodes are being created and added to the word tree.

e) *The generation of the inflection forms* (associated to inflection form nodes). The generation consists on traversing the oriented graph of elementary rules. Because an inflection form can be obtained sometimes from another inflected form, a loop can appear. A checking on this graph assures that there are no cycles that could put the program into an infinite loop.

f) *Syllabification*. If there are previously saved inflected forms, the system tries to attach them to newly generated inflected forms, comparing inflection situations and alphabetic/phonetic text. If the previously saved inflected form is not correct or if there are no previously saved inflected forms, the linguist can generate new syllabification forms.

MKT system can run in debug mode for inflection rules, which may help detecting some errors. During runtime, MKT generates fully detailed HTML files for each basic rule. These reports show basic rules written in GRAALAN language with additional information about all elementary rules, variants, conditions, actions, and a complete log for all lemmas that used them.

After generating inflected forms, MKT adds them to the inflected form LKB. Modifying inflected forms automatic generated by the inflection rules or adding other forms involves generating new basic rules grouped in a new compound which will be referenced by the lemma. The consequences of adjustments made by the linguists are almost transparent for them because MKT automatic determines the minimum number of steps for saving the information and it acts with no user help. If the linguist corrects only the syllabification forms then the system will save only those forms. If the linguist corrects the automated generated inflected forms, then MKT will update inflected forms and inflection rules too.

If the linguist modifies, deletes, or inserts a new inflected form for an inflection situation which has a correspondent inflection situation in one of the active basic rule, for that rule will be generated a new version, specific to the current lemma and according to the changes made by the linguist. If the linguist inserts inflected forms for at least one inflection situation with no correspondent in the active basic rules then a new basic rule will be generates having the "shape" of the AVT resulted from the union of all inflection situation with no correspondent in the active basic rules but with at least one inflected form added by the linguist.

Generating new basic inflection rule involves generating new elementary rules with generated transformations lists and replacing the old elementary rules. The newly created elementary rules have one variant each, they reference the lemma text and they have no associated condition.

The generation of lists with actions for transforming the alphabetic and phonetic text of lemma into alphabetic and phonetic text of inflection form uses the Levenshtein algorithm. This algorithm shows how a minimum list of actions like deleting, inserting and replacing can be obtained to convert one given string into another given string [14].

If there was a new basic rule (generated from scratch or generated as a new version of another basic rule), then the list of all basic rules used by the current lemma will be found into a generated compound rule and the current lemma will reference not the old one but the new compound rule. At the end of work session a special function can be called to clean up the knowledge base. The main purpose of this function is to remove the unused basic and compound rules or the equivalents among them. The results of this function are transparent for the linguist and the function takes care about keeping valid references among all lemmas, compound, basic and elementary rules.

.

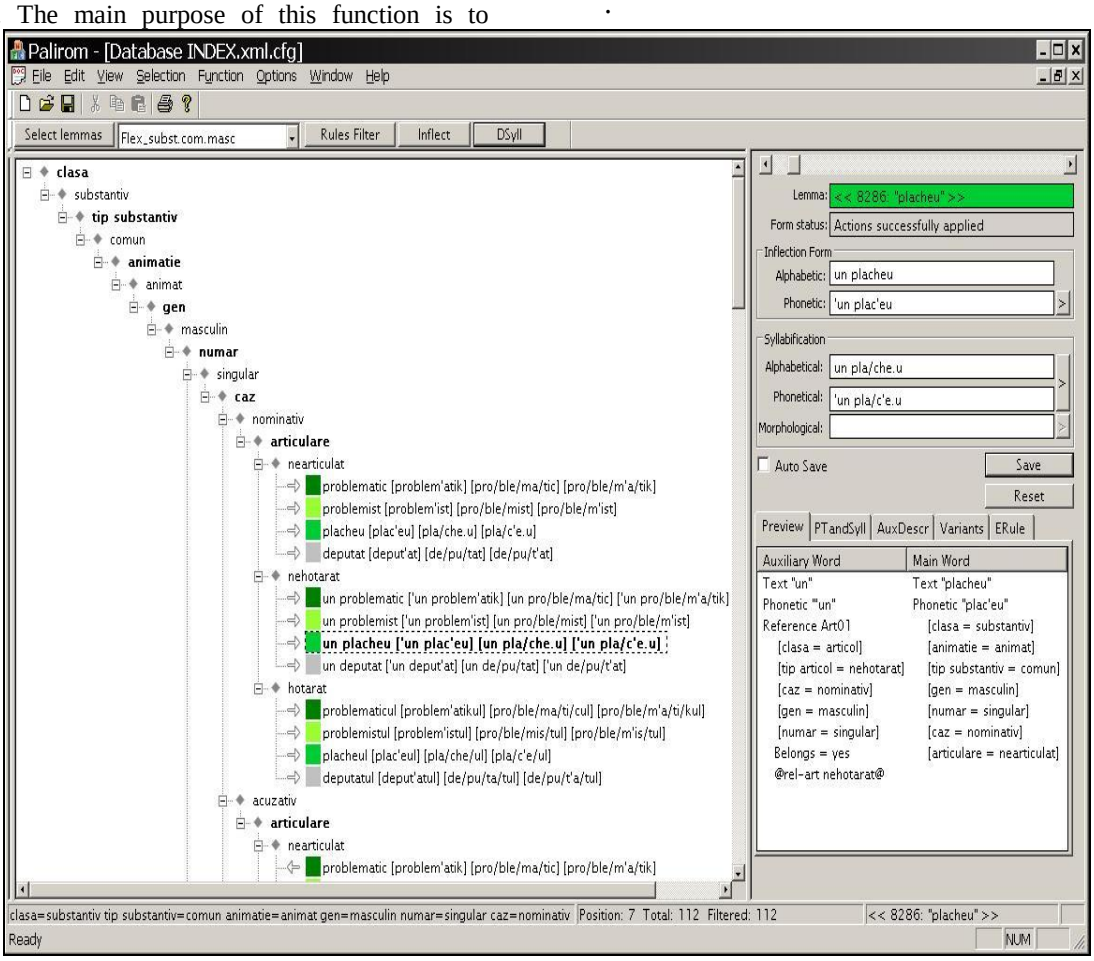

Figure 5. MKT Graphical user interface

X. CONCLUSIONS

The normal and phonetic inflection forms generation process according to GRAALAN language has a great generalisation degree. It can be applied to almost all the natural languages and is particularly efficient for not insulated inflected languages. If the used linguistic knowledge bases (alphabet, morphological configurator, lexicon, syllabification rules, inflection rules) are complete from the inflection process point of view, this process is automatic. Anyway, it is advisable to verify the results in order to see if there are some errors in the input LKB and to see if some exceptions were not omitted, especially in the inflection rules. MKT offers to this purpose a graphical user interface. The set of generated linguistic knowledge bases can be used in other types of application, for example, starting with the simplest ones like spellers, lemmatisers (that use only mono-word inflection forms), mono-word annotators (that use the mono-word inflection forms but also the inflection situations), up to more complex applications like grammar checkers, multiword annotators (that need information from the other GRAALAN knowledge bases too). Finally, if we dispose of the description of many natural languages, the GRAALAN correspondences between pair of languages can be described and, after that, used in automatic or assisted machine translation.

REFERENCES

[1] V. AUBERGE, R. BELRHALI, La phonétisation automatique du français: émergence de règles ou de lexiques, Institut de la communication parlée, Grenoble, FRANCE 1991

[2] A.-M. BARBU, Romanian Lexical Data Bases: Inflected and Syllabic Forms Dictionaries, The 6th edition of the Language Resources and Evaluation Conference, 2008

[3] A. BIDU-VRĂNCEANU, C. CĂLĂRAŞU, L. IONESCU-RUXĂNDOIU, M. MANCAŞ, G. PANĂ-DINDELEGAN, Dicționar de științe ale limbii, Editura Nemira, 2005

[4] C. CHARRAS, T. LECROQ, Sequence comparison, LIR Laboratoire d'Informatique de Rouen and ABISS Atelier Biologie Informatique Statistique Socio linguistique

[5] I. COTEANU, L. SECHE, M. SECHE, et al., DEX Dicționar Explicativ al Limbii Române, Editura Univers Enciclopedic, 1998

[6] S. DIACONESCU, Some Properties of the Attribute Value Trees Used for Linguistic Knowledge Reoresentation, in 2nd Indian International Conference on Artificial Intelligence IICAI-05, India, 2005

[7] S. DIACONESCU, Natural Language Understanding Using Generative Dependency Grammars, in Max Bramer, Alun Preece and Frans Coenen (Eds), ES 2002. Research and Development in Intelligent Systems XIX, Proceedings of ES2002, the Twenty second SGAI International Conference on Knowledge Based Systems and Applied Artificial Intelligence, Cambridge UK, December 2002, Springer, pp.439-452

[8] S. DIACONESCU, Complex Natural Language Processing System Architecture, in Corneliu Burileanu, Horia-Nicolai Teodorescu (Eds.), Advances in Spoken Language Technology, The Publishing House of the Romanian Academy, Bucharest 2007, pp. 228-240

[9] M.Șt. ILINCA, Gramatica Instrumentală, Editura Festina, 1995

[10] International Phonetic Association, Handbook of the International Phonetic Alphabet, Cambridge University Press, 2005

[11] Y. MARCHAND, C. R. ADESTT, R. I. DAMPER, Automatic Syllabification in English: A Comparison of Different Algorithms, 2008

[12] The UNICODE Consortium, The UNICODE Standard, Version 5.0, Fifth Edition, Addison-Wesley Professional, 2006

[13] W3C, Extensible Markup Language (XML) 1.0, Recommandacion, 1998

[14] ***, UNITEX 1.2 User Manual, Université de Marne-la-Vallée, 2006

[15] *** The Penguin English Dictionary, Editura Litera International, (copyright Merriam-Webster Inc.), 2003